Health care and social inference systems: An unauthorized inference control based on fuzzy logic

Souhila Kaci[1], Abdeslam Ali-Laouar[2], and Frédéric Cuppens[2]

[1] Centre de Recherche en Informatique de Lens (C.R.I.L.–C.N.R.S.)
Rue de l'Université SP 16 62307 Lens France

[2] Institut de Recherche en Informatique de Toulouse (I.R.I.T-C.N.R.S)
118 route de Narbonne 62077 Toulouse France

Abstract. In this paper, we address the problem of unauthorized inference of confidential information in the field of health care and social information systems. More precisely, we will focus on the problem of inference control of confidential information from statistical databases which contain information about patients and propopse a method based on fuzzy logic to avoid unauthorized inference. Information provided using our approach remains relevant because it is without loss of quality.

1 Introduction

The security of information systems is a very important problem which has been mainly addressed in military applications. This led to security policies which are applicable only in environments which accept a rigid bluk-heading of information and services handling this information. Indeed, these models cannot be used in other domains which also require security policies like for example the health care domain where it is important to guarantee the confidentiality, integrity and availability of pieces of information contained in medical files of patients. The confidentiality consists in expressing who has the right to reach which information about which, when, and possibly under which conditions. The integrity is the property which ensures that information is modified only by the users authorized under the conditions normally envisaged. Lastly, the availability is the aptitude of an information system for being able to be employed by the users competent under the conditions of accesses and use normally envisaged.

In this paper, we particularly address the problem of security of information systems in the field of health care and social. Let us note that in spite of the development of security policies in this context [6, 7], it is always possible for an external attacker and, especially, for an internal user badly disposed, to try to circumvent the mechanisms of access control to the resources in order to attack the confidentiality, the integrity or the availability of information.

To prevent the infringements against the intimacy of the patients, the medical databases must protect not only confidential information, but also information not explicitly confidential which can be employed to obtain confidential information. This paper treats

Kaci S., Ali-Laouar A. and Cuppens F. (2004).
Health care and social inference systems: An unauthorized inference control based on fuzzy logic.
In *Proceedings of the 2nd International Workshop on Security in Information Systems*, pages 217-226
DOI: 10.5220/0002674902170226

detection and the limitation of the situations for which there is a risk of illegal inference (called also illegitimate inference). This problem is called *unauthorized inference problem*. It can also be simply defined in the following way. Suppose that a user is authorized to access to some information. The crucial question now is: can this user use this information to deduce a confidential information for which she would not have the right of access? A possible solution to this problem is to refuse to answer when this may allow to deduce confidential information however this solution is not interesting because it does not respect the availability condition. Another possible solution is the use of false answers for users having a restricted access to the information system. Indeed this method allows to protect confidential information by providing false but not very significant answers. The problem of this method is that the user to whom one provides false answers can make bad decisions. It is also difficult to provide a coherent set of false answers. The solution that we propose in this paper does not consist to provide a false answer to the user but a *"vague"* information formalized in *fuzzy logic* [8, 4].

Section 2 describes the problem of illegitimate information from databases containing information about the patients. We also describe a well-known method to attack such databases. In section 3, we first present the general principle of our approach. We then give some necessary background on fuzzy logic on which our approach is based. Lastly, section 4 gives a detailed description of our approach.

2 Illegitimate inference in statistical databases

The main difference between a statistical database (SDB for short) and a traditional one relates to the interrogation interface more limited in the SDB. The queries on a SDB are limited to operations like counting (COUNT), sum (SUM), the average (AVG) and other statistical calculus, which are carried out on subsets of data. Although these operations seem to be without consequence, it should be made sure that significant information on the individuals are not revealed. This problem becomes particularly difficult if we accept the possibility that a sequence of general queries, each one by itself does not allow to deduce confidential information, can be employed to deduce significant information. Let us now give an example to illustrate the difficult nature of the inference problem in the statistical databases. We consider a database, given in Table 1, which contains

Table 1. Example of a statistical database.

Name	Sex	Age	Department	salary
Jean	M	27	Mathematics	2.000
Thomas	M	43	computer science	3.000

Name	Sex	Age	Department	salary
Isabelle	F	27	Mathematics	2.600
Justine	F	31	computer science	3.200

information concerning the employees. Let us suppose that the policy of the company imposes that the *salary* of the employees is a confidential information which should not be revealed. To achieve this goal, the database does not return an answer to a query like: *how much is the salary of the employee whose name is Isabelle?* since the answer

is confidential. Similarly, the base does not answer any query when, for example, the average is calculated on the basis of a simple record, i.e. a query concerning only one individual. Consequently, it refuses to answer for example the query: *how much is the average salary of the women employees who work for the computer science department?* because the average here is calculated from only one record.
A query on a SDB R consists to compute a subset of R using a characteristic formula C, which is a logical formula built from the values of the attributes of R by using the logical operators $\wedge$ (and), $\vee$ (or), and $\neg$ (not). For example, the subset of records representing the *women employees who work for the computer science department*, can be represented by the following characteristic formula:

$$C = (\textit{sex=F}) \wedge (\textit{department=computer science}).$$

The set of records which satisfy the characteristic formula C, denoted by X_C, is called the result of the query. Applying the formula C on the relation R given in Table 1, we get: $COUNT(C) = 1$, $AVG(Age, C) = 31$ and $SUM(Salary, C) = 3200$.

Generally, a statistical query taken separately does not allow to deduce confidential information. For this reason, a user with good intentions should be able to form any interesting characteristic formula, and to carry out any statistical measurement on the resulting set of the records. However, it is possible that a user forms statistical queries which can be employed to deduce specific values of a field of the database, which is not acceptable if the values represent confidential information. In this case, we say that the database has been compromised.

A characteristic formula used in order to compromise a database is called a tracker [2, 3]. This formula is chosen so that it gives as a result a set X_C whose size is equal to 1. Denning et col. [2] have shown that for any real database, a tracker can always be found.

In the next section, we propose a new strategy to prevent attacks based on trackers.

3 Our approach

In the everyday life and particularly in the medical field, medical analyses are generally expressed by linguistic descriptions (Example: Temperature of the body is raised, normal, etc). This is especially used for the non-specialists in the medical field. In this paper, we take as a starting point this method to deal with the illegitimate inference problem in statistical databases. More precisely, we replace the results of the statistical queries (quantitative answers) by linguistic descriptions (qualitative answers) in order to limit the risk of illegitimate inference.
For this, our idea consists in replacing the *numerical answers* (e.g. numbers of patients = 10) by *linguistic descriptions* (e.g. medium) formalized in fuzzy logic framework.
Intuitively, each numerical answer is associated to a given class then a qualitative answer is associated to each class. Thus, the formalization of our approach requires two steps: *classification* and *fuzzification*. Let us recall these two concepts:

– **Classification** is the procedure which consists in decomposing the scale of the used numerical values into non-empty classes so that each numerical value belongs to one and only one class.
Let I be a set of elements. We say that $Q(I)$ is a partition of I if there exists a set

$\{q_1, q_2, \cdots, q_k\}$ satisfying the following conditions:

$$\bigcup_{i=1,\cdots,k} q_i = I \text{ with } q_i \neq \emptyset \text{ and } q_i \cap q_j = \emptyset \text{ for } i \neq j.$$

To be relevant, a partition should be made up of definitely individualized classes. Among existing classification methods, we recall one method, that we will use later, based on the aggregation around the centers using a fixed number of classes. The principle of this method is to determine a partition of I composed of k classes, the number k being fixed a priori by the user of the method. k centers $c_1, \cdots, c_k$ are chosen which are either arbitrarily points in the space of the variables, or elements of the set I.

Each element of the set I is associated to one and only one class whose center is one of the k centers $c_1, \cdots, c_k$ according to the following assignment rule:

i belongs to the class q_j of center c_j iff $||i - c_j|| = min_{l=1,\cdots,k}||i - c_l||$.

After the classification step, we have to associate an appropriate linguistic variable to each class. For example, if the numerical scale corresponds to the temperature then the linguistic variable which corresponds to the interval $[20, 25]$ may be *tepid*. This can be formalized in fuzzy logic [8].

– **Fuzzification:** A principal characteristic of the human reasoning is that it is based on vague or incomplete data. Thus, to determine if a temperature is hot or cold is easy for any individual without necessarily knowing its exact value. Fuzzy logic has the aim of studying the representation of vague knowledge and the approximate reasoning. A principal characteristic of fuzzy logic is that an object may belong to a set and at the same time to its complement. Thus, a temperature of 22 may at the same time be hot and not hot.

A *linguistic variable* is a triple (X, V, F_X), where X is a variable (age, temperature, etc) defined on a set of reference V (the set of integers, reals, etc). $F_X = \{A_1, A_2, \cdots\}$ is a finite or infinite set of subsets of V used to characterize X (old, young, hot, cold, etc). Each fuzzy subset represents a linguistic description.

The variable may belong to one or more subsets of this element of reference. For example, the temperature $T = 28$ may belong to the subset "pleasant" but may also belong partly to the subset "hot".

The membership relation between a variable and a subset is called *membership function*. In other terms, we speak about the membership degree of a variable x to a subset F, denoted by $\mu_F(x)$.

A *fuzzy set* F of universe Ω (a fuzzy subset of Ω) is defined by a membership function μ_f which associates to each element x of Ω a value in the interval $[0, 1]$.

$$\mu_F : \Omega \rightarrow [0, 1]$$
$$x \mapsto \mu_F(x)$$

$\mu_F(x)$ represents the membership degree of x to the set F. By definition, if $\mu_F(x) = 0$ then x does not belong to F and more $\mu_F(x)$ approaches 1, more the value x belongs to F. If $\mu_F(x) = 1$ then x belongs completely to F.

A fuzzy subset is said to be convex if and only if:

$$\forall x, y; x > y, \forall z \in [x, y], \mu_F(z) \geq min(\mu_F(x), \mu_F(y)).$$

Generally, we express numerical quantities by vague linguistic descriptions such as "approximately 100". The results of fuzzy measurements or an error analysis are

modelled by fuzzy sets called *fuzzy quantities*. A fuzzy quantity Q is a fuzzy set in the universe $\mathbb{R}$ of real numbers. It is supposed to be normalized.
A *fuzzy interval* N is a convex fuzzy quantity. It is a generalization of a real interval whose extremities are fuzzy in order to model concepts such as "approximately", "roughly", etc.
– Representation of a L-R fuzzy interval A fuzzy interval of type LR has a membership function built from a quadruplet $A = (m_1, m_2, a, b)$, where m_1, m_2, a and b are strictly positive real numbers, and of two functions L and R from $\mathbb{R}^+$ into the interval $[0, 1]$ semi-continuous, non-increasing and satisfying the conditions:

- $L(0) = R(0) = 1$,
- $L(1) = 0$ or $\forall x \in \mathbb{R}^+, L(x) > 0$ and $lim_{x\to+\infty} L(x) = 0$,
- $R(1) = 0$ or $\forall x \in \mathbb{R}^+, R(x) > 0$ and $lim_{x\to+\infty} R(x) = 0$.

The membership function is defined as follows:

$$\mu_F(x) = \begin{cases} L(\frac{m_1 - x}{a}) \text{ if } m_1 - a \leq x \leq m_1 \\ 1 \text{ if } m_1 < x < m_2 \\ R(\frac{x - m_2}{b}) \text{ if } m_2 \leq x \leq m_2 + b \\ 0 \text{ if } x < m_1 - a \text{ or } x > m_2 + b \end{cases}$$

When $m_1 = m_2 = m$, the fuzzy interval $P = (m, m, a, b)_{LR}$ is called *a fuzzy number*, denoted by $P = (m, a, b)_{LR}$ and whose membership function is defined as follows:
$\mu_F(x) = L(\frac{m-x}{a})$if $x < m$, $\mu_F(x) = 1$if $x = m$ and $\mu_F(x) = R(\frac{x-m}{b})$if $x > m$.
Let $P_1 = (p_1, \alpha_1, \beta_1)_{LR}$ and $P_2 = (p_2, \alpha_2, \beta_2)_{LR}$ be two LR-fuzzy numbers. Then the addition $\oplus$, the substraction $\ominus$ and multiplication $\otimes$ are defined by [4]:

$P_1 \oplus P_2 = (p_1 + p_2, \alpha_2 + \alpha_2, \beta_1 + \beta_2)_{LR}$.
$P_1 \ominus P_2 = (p_1 - p_2, \alpha_1 + \alpha_2, \beta_1 + \beta_2)_{LR}$.
$P_1 \otimes P_2 = (p_1 * p_2, p_1 * \alpha_2 + p_2 * \alpha_1, p_1 * \beta_2 + p_2 * \beta_1)_{LR}$.
Contrary to the addition and subtraction, the multiplication $P_1 \otimes P_2$ is not of type LR. An approximate value of type LR is given when P_1 and P_2 have a support included in $\mathbb{R}^+$, α_1 and β_1 are small w.r.t. p_1 and, α_2 and β_2 are small w.r.t. of p_2.

To apply a linguistic representation to a quantitative variable, the principle consists in breaking up all possible values of the given quantitative variable into subsets (a set of classes of values), so that the borders of the classes are not clearly given. This treatment allows to transform a numerical input into a fuzzy subset. The decomposition should not be arbitrary but founded on criteria, such as the homogeneity of the classes, the uniform partition of the universe, the subsets are totally ordered. These subsets are also called *"linguistic variables"*.
The subsets are characterized by their associated membership functions; we associate a membership function to each subset. Their positions and overlappings can be chosen arbitrarily provided that the following conditions are verified: their form should be convex, the subsets (often in the form of trapezoid) should be partially overlapped so that there are no unspecified ranges and lastly to avoid to imbricating more than two subsets.

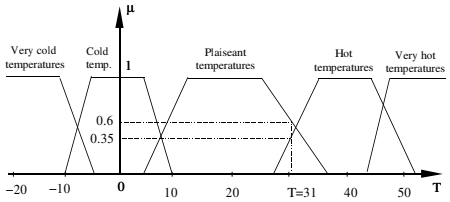

Fig. 1. Representation of the temperature in fuzzy logic.

Example 1. Let us consider the temperature input $T = 31$. According to the membership function given in Figure 1, we obtain the following values:
$\mu_T(\text{very cold temperatures}) = \mu_T(\text{cold temperatutres}) = 0, \mu_T(\text{pleasant temperatures}) = .6$, $\mu_T(\text{hot temperatures}) = .35$ and $\mu_T(\text{very hot temperatures}) = 0$.

Now, it seems important to answer some questions : How many *classes* is it necessary to represent each quantitative variable? Which are the best *linguistic values* for each class? For the first question, more the number of linguistic values is high, more the partitioning quality is good. It is necessary however that the rate: $Card(\Omega)/$*Number of Partitions* is not equal to 1, otherwise this simply means that there is no fuzzification.
For the second question, we compute the membership degree of each element x to all the subsets F_i of the universe Ω. Let $\mu_{F_i}(x)$ be the membership degree of x to F_i. We say that $x \in F_i$ only if $\forall F \in \Omega, \mu_F(x) \leq \mu_{F_i}(x)$.

4 Detailed description of our approach

The principle of our method consists, in a first step, to decompose the set of values of the confidential attributes into subclasses of values. Each subclass contains values according to a given criterion. In this paper, we will use the classification method based on a fixed number of classes.
After the classification into subclasses the fuzzification comes. We transform each class into a fuzzy quantity i.e., a fuzzy number with a membership function. Then, we associate a linguistic variable to each number (small, large, etc). Next, for each answer provided by the database management system, we compute the membership degree of this answer to each fuzzy subset (linguistic variables). The answer of our system is the linguistic variable which has the highest membership degree.
Let us note that the simplest version of a statistical query SQL is written as follows:

SELECT f(<attributes>) *FROM* <relations> *WHERE* <conditions>,

where f is a statistical function such as Sum, Avg, $Count$, etc.

In this paper, we focus on queries which compute *statistical quantities*, i.e. queries which deduce information on aggregation such as sum, $average$, max and min.
Let us consider the example of relation R (patient, H/F, age, sickness insurance company, leucocyte rate) given in the Table 2 (borrowed from [5]).
The number of patients is 10 and the normal leucocyte rate in mm3 of blood is 4500.
In this example, we suppose that the *leucocyte rate* is a confidential attribute. To control the illegitimate inference on this attribute, we will transform the answers to the queries

Table 2. Example of a database.

Patient	M/F	Age	Sick. ins.	Leucocyte	Patient	M/F	Age	Sick. ins.	Leucocyte
Dufour	M	45	MAAF	4000	Dupont	M	30	MMA	6000
Dulac	F	35	MMA	7000	Dupr	F	32	IPECA	7200
Dulon	M	55	MGEN	3500	Dupuis	F	50	MGEN	6800
Dumas	M	40	Rempart	3800	Durand	F	25	LMDE	3000
Dumont	M	38	MMA	7500	Duval	M	45	IPECA	5500

concerning this attribute by giving qualitative answers.
We proceed in the same way for the answers to the queries which compute the number of patients who verify a given condition. For this, we fuzzify the number of patients and the leucocyte rate.
Let us start with the number of patients and decompose this variable as follows: A first class: from 0 to 3, a second class: from 4 to 6 and a third class: from 7 to 10.
We now transform each class into a fuzzy number $A_i(m, a, b)$ where m is the center of the class, a and b represent the degrees of inaccuracy.
For each number, we associate a linguistic variable (see also Figure 2-a):

– The first class is fuzzified by the fuzzy number "$small$" $= (2, 2, 2)_{LR}$,
– The second class is fuzzified by the fuzzy number "$medium$" $= (5, 2, 2)_{LR}$,
– The third class is fuzzified by the fuzzy number "$great$" $= (8, 2, 2)_{LR}$.

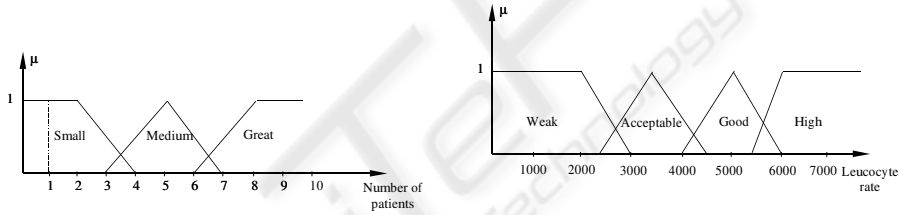

Fig. 2. (a) Fuzzification of the number of patients. **(b)** Fuzzification of the leucocyte rate.

We now classify the leucocyte rate for a patient as follows:

– 1^{st} class: from 0 to 3000, 2^{nd} class: from 3000 to 4500,
– 3^{rd} class: from 4500 to 6000 and 4^{th} class: from 6000 to 7000.

We now propose the following fuzzification (see also Figure 2-b):

– The first class is fuzzified by the fuzzy number "$weak$" $= (2000, 1000, 1000)_{LR}$
– The second class is fuzzified by the fuzzy number "$acceptable$" $= (3500, 1000, 1000)_{LR}$
– The third class is fuzzified by the fuzzy number "$good$" $= (5000, 1000, 1000)_{LR}$
– The fourth class is fuzzified by the fuzzy number "$high$" $= (6000, 1000, 1000)_{LR}$

Let us now suppose that a user is authorized to carry out statistical queries and she wants to discover the leucocyte rate of "Dulon". Let us also suppose that this user knows moreover that "Dulon" has the MGEN as a sickness insurance company. Consider now the following queries:

1) SELECT Count(Patient) FROM R WHERE M/F='M' AND Sick. ins. ='MGEN'

$Result = 1$ (R_1)

Let us compute the membership degrees $\mu_{F_i}(R_1)$ of the result (R_1) w.r.t. each fuzzy subset. We get: $\mu_{small}(R_1) = 1$, $\mu_{medium}(R_1) = 0$ and $\mu_{great}(R_1) = 0$.
So the answer provided after fuzzification is "small" since it corresponds to the highest membership degree.

2) SELECT AVG(Leucocyte) FROM R WHERE M/F='M' AND Sick. ins. = 'MGEN'

$Result = 3500$ (R_2)

We compute the membership degrees $\mu_{F_i}(R_2)$ of the result (R_2) w.r.t. each fuzzy subset: $\mu_{weak}(R_2) = \mu_{good}(R_2) = \mu_{high}(R_2) = 0$ and $\mu_{acceptable}(R_2) = 1$.
Then the answer is "acceptable".
Note that the deduction of confidential information when handling numerical answers is very easy. It is clear that from (R_1) and (R_2), the user may directly deduce that the leucocyte rate of "Dulon" is equal to 3500.
The case of the qualitative answers is less simple: from (R_1), the user knows that the size of the set to which "Dulon" belongs is "small", and from (R_2), she deduces that their average of the leucocyte rate (the set "small") is "acceptable".
Let us now see what may the user deduce from these two information. For this, we know that the average is defined by the equation $\overline{x} = \frac{1}{n}\sum x_i$. It is clear that when n is equal to 1, the average is equal to x_i. To see the impact of the fuzzification on the reasoning of the user, we will analyze the use of the fuzzification step by step:

- Let us suppose that the number of patients is not fuzzified whereas the leucocyte rate is. The answer given to the user in this case is then: the number of patients is equal to 1 (as an answer to the query (R_1)) and their average leucocyte rate is "acceptable", which allows to deduce that the leucocyte rate of Dulon is "acceptable". However, the fuzzification of the number of patients makes that the answer provided to the user (also as an answer to the query (R_1)) is "small", which does not allow to know precisely how many patients correspond to this answer "small".
- Let us now suppose that the user knows moreover that the maximum size of the fuzzy quantity "small" is equal for example to two. However even if the user has this information, she deduces nothing as we will show on the following example: It is known that an "acceptable" leucocyte rate lies between 3000 and 4500. Let the size of "small" be equal to 2. From a leuycocyte average of two patients x_1 and x_2 equal to "acceptable", we may have the following possibilities for x_1 and x_2:
  - $x_1 = 2500 \equiv$ "weak"[3], $x_2 = 4000 \equiv$ "acceptable"
  - $x_1 = 2500 \equiv$ "weak", $x_2 = 5000 \equiv$ "good"
  - $x_1 = 1500 \equiv$ "weak", $x_2 = 6500 \equiv$ "high"
  - $x_1 = 3500 \equiv$ "acceptable", $x_2 = 5000 \equiv$ "good"
  - $x_1 = 3500 \equiv$ "acceptable", $x_2 = 4000 \equiv$ "acceptable".

  From these results, one can say that from a leucocyte average equal to "acceptable" computed for two patients, one concludes nothing on the leucocyte rate of one of the two patients.

[3] The equivalence means here that the number (e.g. 2500) corresponds to the given class (e.g. "weak") after fuzzification.

Note that to have an average rate "acceptable", we have five possibilities for the leucocyte rate for each of the two patients. In only one case, the rate of the two patients is "acceptable". In the other cases, it varies between "Weak", "acceptable", "Good" and "High". So we have four cases with x_1 or x_2 equal to "acceptable" and six cases different from "acceptable".
Then we may say that it is totally possible that the leucocyte rate of "Dulon" is equal to "acceptable", but it is also totally possible that it is different from "acceptable". Indeed, we are in a situation of total ignorance.

Let us note that in the real case, the database may contain thousands of patients and the fuzzy quantity "small" may reach several hundreds of patients. Consequently, the possibilities of inference are even weaker when the cardinality which corresponds to the fuzzy quantity is larger. The user deduces nothing on the leucocyte rate of "Dulon" when all the possible cases are considered.

3) SELECT Count(Patient) FROM R. Then, $Result = 10$ (R_3)
The answer after fuzzification is "great" (R'_3)
The user tries thereafter to know the number of patients different from "Dulon". For this, she gives the following query:

4) SELECT Count(Patient) FROM R WHERE NOT (M/F='M' AND Sick. ins.='MGEN');
$Result = 9$ (R_4)
The answer after fuzzification is "great" (R'_4)
From these two answers, the user may construct the following reasoning: The difference between (R_3) and (R_4) (10-9=1) corresponds to the number of male patients who have the MGEN as a sickness insurance company (i.e., the number of patients having the same properties as "Dulon").
With a similar reasoning, she concludes that the difference between (R'_3) and (R'_4) is equal to [4] $\mid "great" \ominus "great" \mid = \mid (8,2,2)_{LR} \ominus (8,2,2)_{LR} \mid = \mid (8,2,2)_{LR} \oplus (-8,2,2)_{LR} \mid = \mid (0,4,4)_{LR} \mid$ which is equivalent to $(0,0,4)_{LR}$ after removing the negative part, since there is no negative leucocyte rate.
So we have $(R'_3) \ominus (R'_4) \sim "small"$. Indeed, we get the same result as for (R_1) after fuzzification.
To know the average of the leucocyte rate for all the patients, the user gives the following query:

5) SELECT AVG(Leucocyte) FROM R. Then, $Result = 5430$ (R_5)
The answer after fuzzification is "good" (R'_5)
To compute the average of the leucocyte rate of all the patients different from "Dulon", the user gives the following query:

6) SELECT AVG(Leucocyte) FROM R WHERE NOT (M/F='M' AND Sick. ins.='MGEN'),
$Result = 5644$ (R_6)
The answer after fuzzification is "high" (R'_6)
In the case of numerical answers, to know the leucocyte rate of "Dulon", the user computes the following value: $10 * 5430 - 9 * 5644 = 3500$.
With a similar reasoning, in the case of qualitative answers, she may try to proceed

[4] Since the values are not known a priori but supposed to be positive, the subtraction is translated into fuzzy logic by the absolute value.

in the following way. The leucocyte rate of "Dulon" is equal to:
$| ((R'_3) \otimes (R'_5)) \ominus ((R'_4) \otimes (R'_6)) | \sim | (-8000, 38000, 38000)_{LR} |$.
From the obtained number, the user deduces nothing because the leucocyte rate is never negative. Even if she can deduce some information (if the fuzzification is changed), the situation is similar to the first case since the user does not know the exact number of patients. Let us also note that we lost the precision on the computation of the leucocyte rate because of the multiplication which we carried out (recall that in the case of the multiplication, the computation is only approximate).

We have shown on this example that the user may use different ways to deduce confidential information however the use of qualitative answers makes difficult the implementation of attacks by trackers because after fuzzification, it is difficult to identify the individual concerned by the confidential information. Indeed, required information is not distinguished after fuzzification.

5 Conclusion

We have proposed a first attempt to limit the risk of inference of confidential information from a database using fuzzy logic. It is difficult to affirm here that we eliminate any risk of illegal inference. The goal is nevertheless to continue to answer the queries as well as possible using non-confidential information. So our aim is to limit at least as possible the restrictions of legitimate access on databases while ensuring that the risk of unauthorized inference remains below an acceptable threshold.

An immediate prospect for this work would be to implement our approach and to validate it on great databases. We showed in this paper that our approach particularly enables us to control the attacks by trackers. We expect to see how this approach could be used to control other types of attacks like linear systems [2, 1]. Lastly, it would be interesting to see to what extent our approach is sensitive to the classification method used, i.e. to see if the use of other classification methods give sensitively different results.

References

1. F. Cuppens. A logical analysis of authorized and prohibited information flows. In IEEE Symposium on Research in Security and Privacy, 1993.
2. D. Denning, P. Denning and Schwartz. The tracker: A Threat to Statistical Database Security. ACM Transactions on Database Systems, 4(1): 76-96, 1979.
3. D. Denning and J. Schlorer. A Fast Algorithm for Calculating a Tracker in Statistical Databse. ACM Transactions on Database Systems, 5(1), 1980.
4. D. Dubois and H. Prade. La logique floue. In Quaderni, 50-73, 1995.
5. A.A. El Kalam. MP6, Sous-projet 3: Politiques de scurit pour les SICSS. Informations protger et menaces. Rapport technique.
6. R. Sandhu, E. Coyne, H. Feinstein and C. Youman. Role-based access control models. IEEE Computer, 29:38-47, 1996.
7. S. Solms. The management of computer security profiles using a role-oriented approach. Computer and Security, 13(8), 673-680, 1994.
8. L. Zadeh. Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets and Systems, 1, 3-28, 1978.